ECM - Capture

Dr. Ulrich Kampffmeyer

PROJECT CONSULT
Unternehmensberatung
Dr. Ulrich Kampffmeyer GmbH

Hamburg, 2009

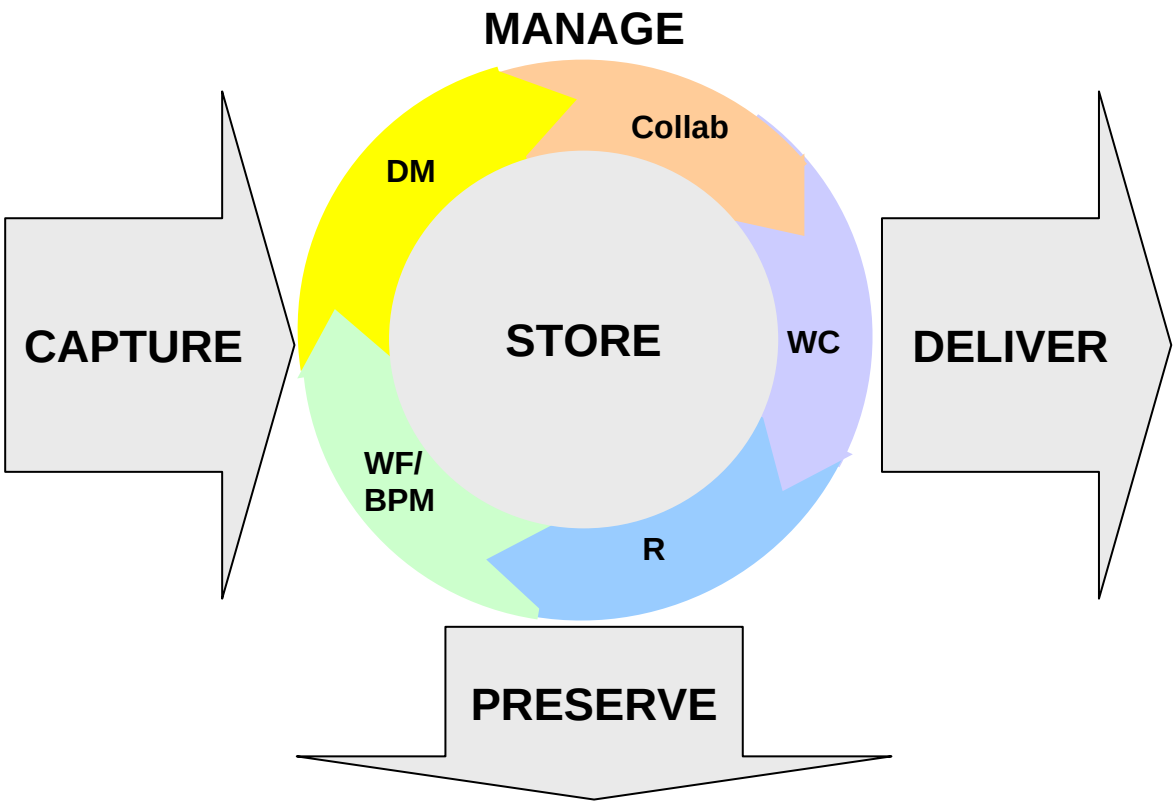
Einleitung

Dokumentenmanagement- oder Enterprise-Content-Management-Systeme (folgend: ECM) ganz gleich welcher Colour ziehen ihren Nutzen immer aus den bereitgestellten Informationen. Aber bevor diese genutzt werden können, müssen sie den entsprechenden Systemen auch zugeführt werden. Deswegen ist die erste und wichtigste Stufe eines ECM-Systems diese Funktionalität. Man verwendet hierbei den Begriff „Capture“.

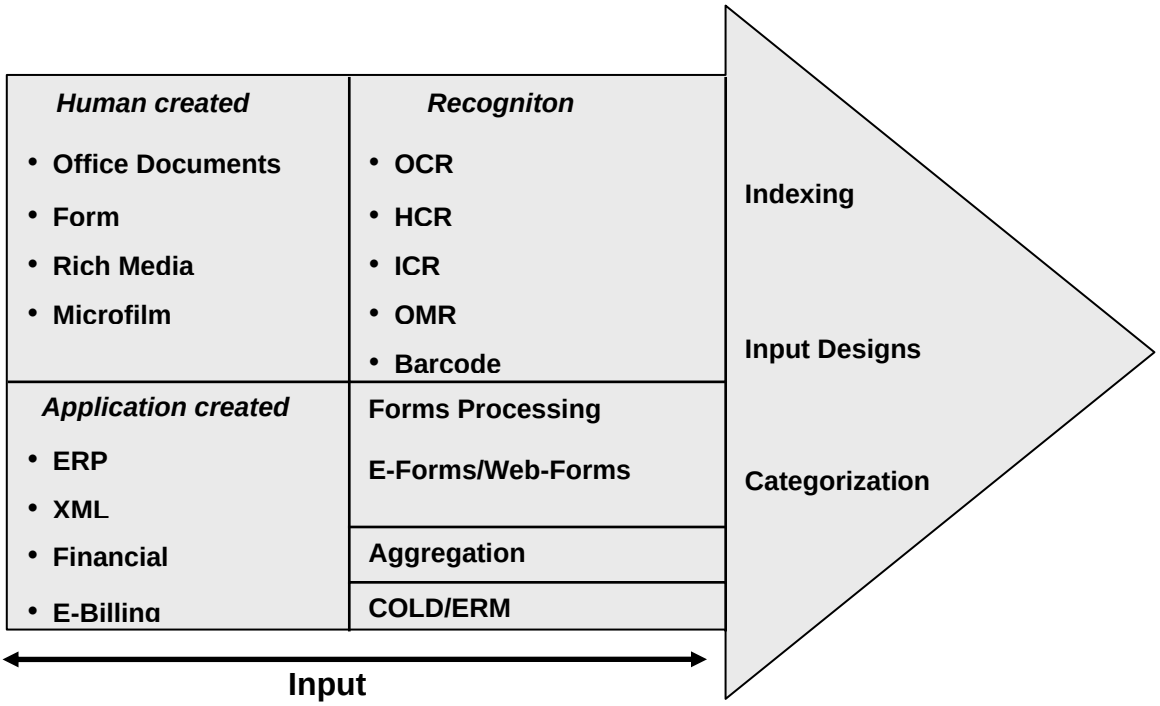
Unter „Capture“ versteht man alle im Zusammenhang mit der Erfassung, Erkennung und Klassifizierung von Dokumenten oder besser Objekten durchzuführenden Tätigkeiten.
„Capture“ hat sich zwischenzeitlich zu einem eigenständigen Bereich entwickelt, der sowohl ECM-Systeme und Archive als auch operative Systeme (z.B. ERP-Anwendungen) mit Daten beliefert. Die Erfassung der im ECM zu verwaltenden Daten ist ein wichtiger Bestandteil beim Dokumenten-Management. Da hierbei zu einem wesentlichen Teil die späteren Nutzungsmöglichkeiten bestimmt werden, sollte man der Erfassung sowohl der Planung als auch später der Durchführung und der Kontrolle entsprechende Aufmerksamkeit widmen. Die unterschiedlichen Dokumentenquellen - Papier, COLD, E-Mail, Office-Dokumente usw. - erfordern in der Regel auch unterschiedliche Erfassungs-, Attributierungs-/ Klassifizierungs- und Prüfschritte. Teilweise kommen bei der Erfassung zusätzliche rechtliche Anforderungen hinzu - etwa bei eingehenden elektronischen Rechnungen oder Belegen im Sozialversicherungsbereich.
Das Spektrum der Erfassungsverfahren ist entsprechend groß. Die verschiedenen Verfahren verlangen teilweise nach recht unterschiedlichen Techniken. In vielen Unternehmen kommen dabei gleich mehrere Erfassungsverfahren parallel zum Einsatz, um dem Anspruch des Enterprise-Dokument-Management-Anspruchs zu genügen und alle für das Unternehmen relevanten Dokumente möglichst integriert und weitgehend vollständig elektronisch zu erfassen und zu verwalten.

Einordnung in das ECM-Modell: Überblick über die Komponenten von „Capture“

Die AIIM (Association for Image and Information Management), eine US-amerikanische Anwender- und Anbietervereinigung, hat bereits vor vielen Jahren ein ECM-Model vorgestellt, welches sehr transparent und anschaulich die Teile einer DMS/ECM-Anwendung darstellt.



Die Kategorie „Capture“ beinhaltet Funktionalität und Komponenten zur Erstellung, Erfassung, Aufbereitung und Verarbeitung von analogen und elektronischen Informationen. Es werden mehrere Stufen und Techniken unterschieden - von der einfachen Erfassung der Information bis zur komplexen Aufbereitung durch eine automatische Klassifikation. Die Capture-Komponenten werden auch häufig als „Input“-Komponenten zusammengefasst und als "Input-Management" bezeichnet.



Generell kann unterschieden werden zwischen Indexing, Input Designs und Categorization.

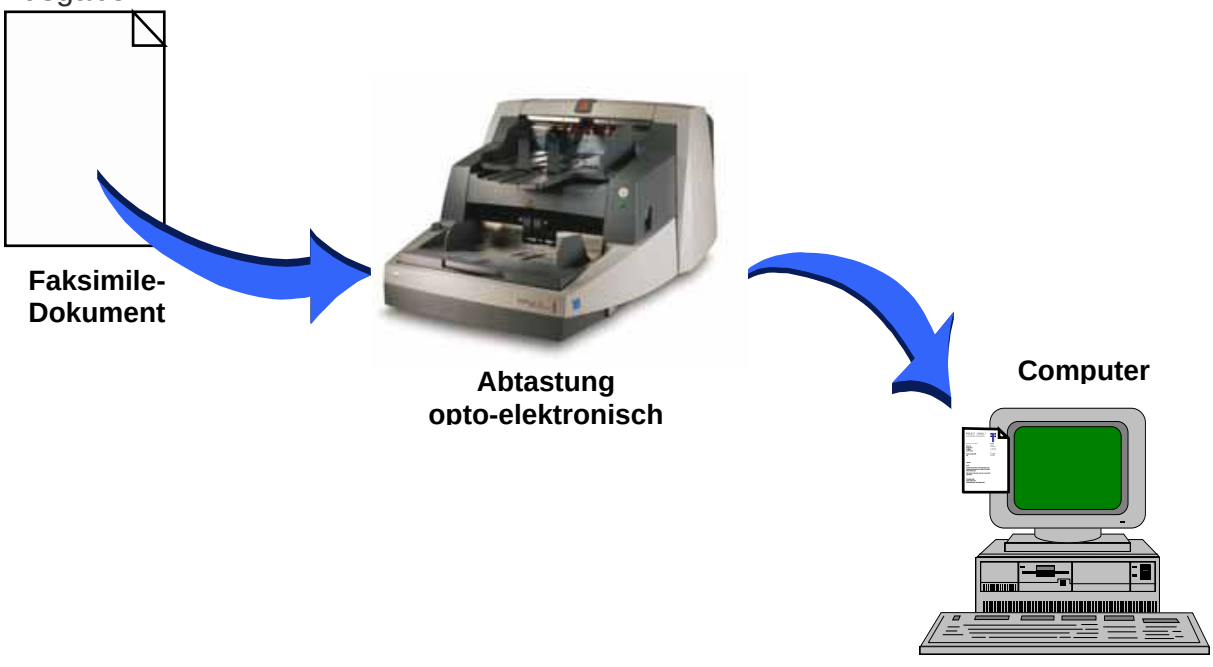
Anders als im Deutschen beschränkt sich im Angloamerikanischen der Begriff „Indexing“ auf die manuelle Vergabe von Indexattributen, die in der Datenbank einer „Manage“-Komponente für Verwaltung und Zugriff auf die Informationen benutzt wird. Im Deutschen werden hier auch Begriffe wie „Indizieren“, „Attributieren“ oder „Verschlagworten“ benutzt.
Sowohl die automatische als auch die manuelle Indizierung kann durch hinterlegte Input Designs (Profile) erleichtert und verbessert werden. Solche Profile können z.B. Dokumentenklassen beschreiben, die die Anzahl der möglichen Indexwerte beschränken oder bestimmte Kriterien automatisch vergeben. Input Designs schließt auch die Eingabemasken und deren Logik bei der manuellen Indizierung ein.

„Categorization“ beschreibt den Prozess der automatischen Klassifikation oder Kategorisierung auf Basis der in den elektronischen Informationsobjekten enthaltenen Informationen (z.B. OCR-gewandelte Faksimiles, Office-Dateien oder Ausgabedateien). Hierbei können Programme zur automatischen Klassifikation selbstständig Index-, Zuordnungs- und Weiterleitungsdaten extrahieren. Solche Systeme können auf Basis vordefinierter Kriterien - oder selbstlernend - Informationen auswerten.
Der Flaschenhals der digitalen Informationsverarbeitung ist vor allem die schnelle Erfassung der Informationen. Sie gilt im besonderen Maße für existierendes Schriftgut, das mittels Scannertechnologie in ein elektronisches Informationssystem überführt werden soll. Dazu zählen Posteingang, sonstige Papierdokumente, eingehende Vordrucke etc. Ein weiteres Problem liegt darin, diese NCI-Dokumente (NCI - Non Coded Information) mit Zugriffsinformationen zu versehen. Dies kann manuell beim Scannen, durch automatisches Erkennen von Text oder Barcode und durch Ergänzung fehlender Informationen aus bestehenden DV-Systemen geschehen. Für die

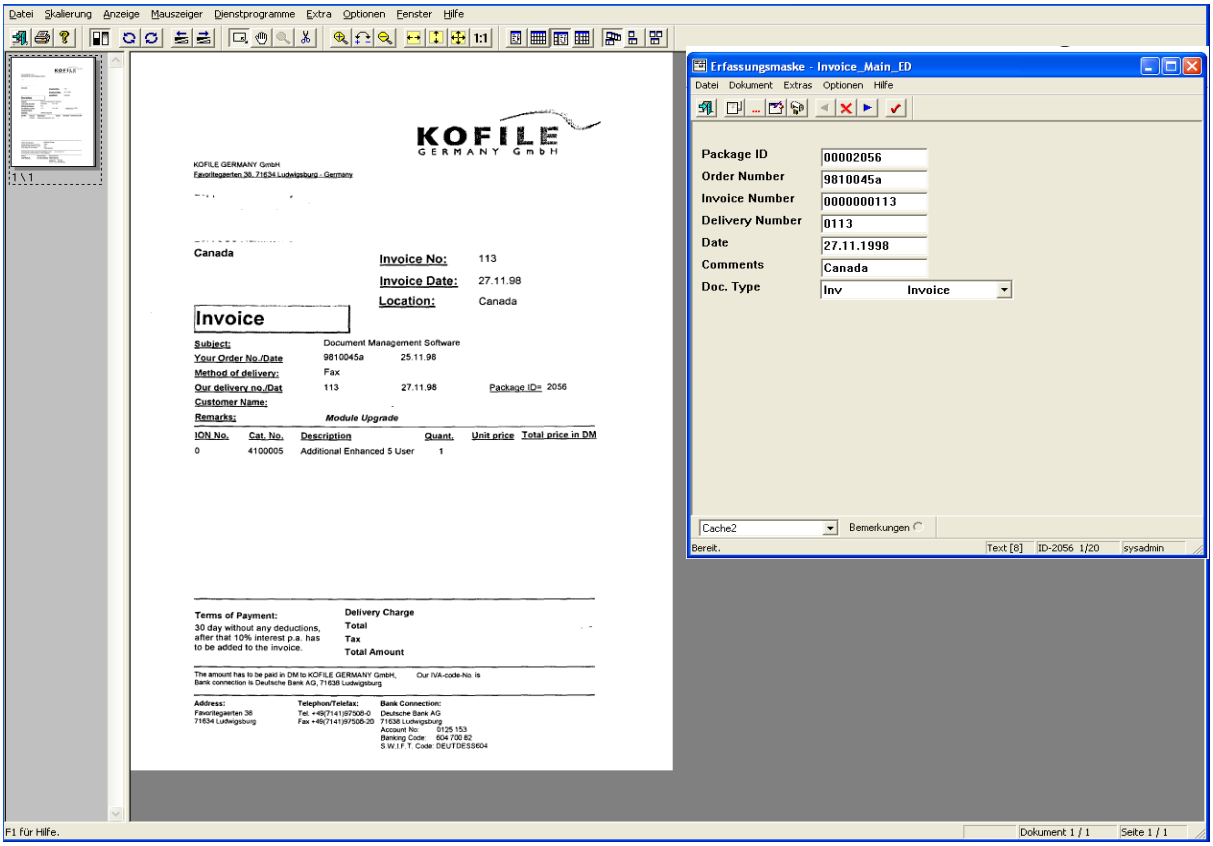
automatische Extraktion von Zugriffsinformationen, der sogenannten Indexinformation, sind Techniken wie OCR (Optical Character Recognition), ICR (Intelligent Character Recognition), HCR (Handprint Character Recognition), OMR (Optical Mark Recognition), Barcode u.ä. erforderlich. Diese können nur unter bestimmten qualitativen Voraussetzungen der Dokumentenvorlagen sicher gewonnen werden. In diesem Rahmen ist auch „Schriftgut“ zu sehen, das gar nicht mehr in Papierform erzeugt wird.
Elektronische Dokumente erlauben eine einfache automatische Indizierung und gewinnen bei der Speicherung von eigen erstellten Dokumenten (z.B. Ausgangsrechnungen, Office-Dokumenten, E-Mails) oder elektronischen, eingehenden Informationsobjekten (z.B. E-Mails, EDI-Dokumente etc.) immer mehr an Bedeutung. Die automatische Übernahme von Daten erfolgt in der Regel im sogenannten COLD-Verfahren.
Um die generelle Lesbarkeit eingescannter Informationen und die Basis für eine optimale Erkennbarkeit für die eingesetzten Extraktions-Techniken sicher zu stellen, ist es ratsam, entsprechende Bildbearbeitungstechniken einzusetzen.
Eine zusätzliche Optimierung des Erfassungsprozesses lässt sich durch die Verarbeitung von Formularen und Vordrucken erreichen. Hierbei werden industriell oder individuell gedruckte Vordrucke beim Scannen erfasst. Zusätzlich kommen anschließend häufig Erkennungstechniken zum Einsatz, da gut gestaltete Vordrucke eine weitgehend automatische Verarbeitung ermöglichen.
Bei der Verarbeitung elektronischer Formulare (E-Forms / Web-Forms) ist eine automatische Erfassung möglich, wenn Layout, Struktur, Logik und Inhalte dem Erfassungssystem bekannt sind.

Manuelle Erfassung: Scannen, Import von Office- und anderen Dateien, Indizieren, Fehlervermeidung beim Indizieren

Die Erfassung von papiergebundenen Dokumenten bezeichnet man als Scannen.Die Begriffe „Scanner“ und „Scannen“ leiten sich von dem englischen Begriff für „Abtasten“ ab. Scannen ist ein Zusammenspiel der Komponenten Scan-Eingabe, Verarbeitung und Ausgabe.



Die Entscheidung, welche Form des Scannens gewählt wird, ist abhängig von den Anforderungen an die Qualität und den Prozess der Erfassung.
Die Einrichtung einer eigenen zentralen Scanstelle (z.B. in der Poststelle) kann dann in Betracht kommen, wenn relativ große Mengen an Dokumenten eingehen und diese taggleich elektronisch am betreffenden Sachbearbeiterplatz zur Verfügung gestellt werden sollen.
Scannen direkt am Arbeitsplatz oder abteilungsweises Scannen wird häufig in Anwendungsfällen mit kleinerem Volumen oder verteilt anfallendem Schriftguteingang eingesetzt. Als Sonderfall ist in diese Kategorie das dezentrale Scannen an entfernten Arbeitsplätzen und anschließende Übertragen der Daten per Leitung (z.B. von Baustellen) zuzuordnen, wobei diese Variante zusätzlich die vorgenannten noch ergänzen kann.
In vielen Fällen kann bei großen Belegmengen oder spezialisierten Aufgabenstellungen (z.B. bei der Indizierung und Klassifizierung) die exklusive oder zusätzliche Inanspruchnahme von Scan-Dienstleistungen (Outsourcing) bei einem entsprechenden Dienstleistungsunternehmen stattfinden.
Es gibt zwei allgemeine Varianten beim manuellen Indizieren: die Erfassung vom Beleg, das sog. Einzelblatt-Scannen oder die Erfassung vom Bildschirm, das sog. Stapelscannen mit anschließendem Indizieren. Im ersten Fall werden die zu erfassenden Indexkriterien vom Beleg entnommen und in einer Erfassungsmaske eingegeben. Erst danach wird der Beleg gescannt und mit den erfassten Daten archiviert. Beim zweiten Fall werden die Belege stapelweise eingescannt und dann ebenfalls in einer Erfassungsmaske sukzessive verschlagwortet (siehe folgende Abbildung).



Die folgenden Varianten finden sich in anwendungsspezifischen Abläufen wider.

Beim Scannen nach der Bearbeitung werden die Geschäftsvorfälle vom Papierbeleg bearbeitet. Je nach Anforderung wird das Dokument danach

- als Papier archiviert,
- nach Abschluss des Vorgangs gescannt, archiviert und vernichtet oder zusätzlich als Original archiviert,
- nach Erstbearbeitung gescannt und vernichtet oder zusätzlich als Original archiviert,
- nach Bearbeitung im Original zurückgesendet und als Kopie archiviert.

Bei Scannen nach der Erstbearbeitung erfolgt die weitere Bearbeitung ausschließlich am Bildschirm.
Beim Scannen vor der Bearbeitung werden die Belege des Geschäftsvorfalls am Bildschirm bearbeitet. Die Indizierung erfolgt entweder vor dem Scannen am Papier oder nach dem Scannen am Image auf dem Bildschirm. In den meisten Fällen nimmt die Indizierung beim Scan-Prozess nur eine erste Klassifizierung vor, die über die weitere Verarbeitung entscheidet. Während der Bearbeitung wird die endgültige Indexzuordnung ergänzt. Die Dokumente werden nach dem Scannen vernichtet oder zusätzlich im Original archiviert.

Das manuelle Indizieren kann auf Grund von Eingabefehlern des Erfassungspersonals ohne entsprechende Maßnahmen und technische Hilfsmittel zu einer mangelhaften Archivierung führen. Deswegen sollte immer dann wenn es möglich ist, ein entsprechendes Verfahren oder eine Kombination davon zum Einsatz kommen.

Vorbelegte Auswahl-Menues zwingen zur Erfassung einzelner Felder von nur voreingestellten Inhalten. Bestimmte Feldtypen erlauben die automatische Befüllung durch die Erfassungssoftware (z.B. Erfassungsdatum = Systemdatum). Mit Hilfe von Prüfziffern können Eingabefehler erkannt werden. Bei Verwendung der Ziffern 0-9 tritt eine zufällige Übereinstimmung auch bei ungültigen Zahlen mit ca. 10°% Wahrscheinlichkeit auf. Bei Verwendung von zwei Prüfziffern liegt diese Fehlerwahrscheinlichkeit nur noch bei ca. 1 %. Die Prüfung von Plausibilitäten trägt ebenso erheblich zur Vermeidung von Erfassungsfehlern bei. So darf z.B. ein Rechnungsdatum nicht nach dem Erfassungsdatum (Systemdatum) liegen. Als besonders sichere Methode ist in diesem Zusammenhang der Abgleich der erfassten Attribute mit bestehenden Datenbank-Inhalten zu nennen. Über die eigentliche Prüfung hinaus kann damit eine automatische Ergänzung weiterer Attribute erfolgen.

Eine wichtige Anforderung innerhalb des Scan-Vorgangs ist die revisionssichere Protokollierung. Diese ist erforderlich, um z.B. die Nachvollziehbarkeit im Sinne der GoBS zu wahren. Die revisionssichere Protokollierung führt in Verbindung mit der Verfahrensdokumentation zu einem sogenannten „elektronischen Dokument hoher Qualität“. Protokollsätze sollen die Angaben von Benutzer, Signaturcode, Datum/Uhrzeit, Unique Identifier des Informationsobjekts, etc. enthalten.

Die Archivierung von Office- und anderen Dateien lässt sich natürlich in vielen Fällen automatisch durchführen. Diese Vorgehensweise wird an anderer Stelle behandelt. Oftmals geschieht das aber auch am Arbeitsplatz des Sachbearbeiters, der diese

Dokumente erstellt und entscheidet, welche archiviert werden sollen. Hierzu werden meistens Drittanwendungen in die Office-Anwendungen eingeklinkt, die die erstellten Dokumente zusätzlich zum Ausdruck an einen „Virtuellen Printer“ schicken. Dabei werden die Indexkriterien automatisch nach bestimmten Vorgaben ausgelesen, das Dokument TIFF (S/W) oder JPEG (Farbe) konvertiert und gleichzeitig archiviert.

Scanner-Technologien, Scanner-Typen, Erfassungsverfahren, Multifunktionsgeräte, Auflösung, Farbe, Qualitätsanforderungen

Bei der Wahl des Scanners kommt es darauf an, welche Art von Dokumenten einzuscannen sind - Texte, Belege, Formulare, Fotos, Dias und vieles mehr. Für jede dieser Vorlagen gibt es Scanner, die sich in Hardwaretechnologie, Software und Komplexität der Bedienung unterscheiden. Je nach notwendigen Weiterverarbeitungs- und anschließenden Ausgabemöglichkeiten unterscheiden sich die Anforderungen des Anwenders.

Mittlerweile sind die überwiegende Mehrzahl der ausgelieferten Dokumentenscanner Duplexgeräte, selbst wenn in der Regel nur die Blattvorderseite zu erfassen ist. Die Preisdifferenz zwischen den Simplex- und Duplexscannern ist inzwischen so gering, dass viele Hersteller Simplexscanner gar nicht mehr anbieten. Ähnlich verhält es sich beim Papierformat: Dokumentenscanner gibt es im A4- und A3-Format, doch vor allem im Marktsegment der Produktionsscanner sind mittlerweile überwiegend DIN-A3-Modelle zu finden. Selbst wenn als Beleggut meistens A4 vorkommt, kann durch quer eingelegte Blätter und automatische Rotation um 90 Grad eine weitere Produktivitätssteigerung erzielt werden. Ein im DIN-A3-Scanner quer (landscape) eingelegtes A4-Blatt wird etwa 30 Prozent schneller als das hochformatig (portrait) zugeführte Dokument verarbeitet.

Vor Einsatz eines Scanners müssen die zu scannenden Belege sorgfältig analysiert werden, um das passende Gerät zu finden: Nicht alle Geräte können Sonderformate, Überlängen oder Endlosbelege scannen. Und nicht jeder Scanner ist in der Lage, Kleinstbelege zu verarbeiten. Sobald Spezialpapier (z.B. NCR-Papier) gescannt werden soll, ist Vorsicht geboten: Nicht alle Einzugsrollen können dies verarbeiten. Auch die Fähigkeit, unterschiedliche Papierstärken zu verarbeiten, ist bei jedem Scanner unterschiedlich. Wenn man die Minimalspezifikationen nicht beachtet, muss mit Doppeleinzügen bei zu dünnem Papier gerechnet werden. Wenn neben Einzelblättern auch gebundene oder empfindliche Dokumente zu bearbeiten sind, ist ein Flachbettscanner mit integriertem ADF die richtige Wahl.

Abhängig von ihrer Geschwindigkeit werden die Dokumentenscanner eingeteilt in: Arbeitsplatz-Scanner (bis 1.000 Belege/Tag), Abteilungs-Scanner (bis 3.000 Belege/Tag), Low-Volume-Production-, Mid-Volume-Production- und High-Volume-Production-Scanner (bis 60.000 Belege/Tag). Es gibt Geräte, die über 200ppm (ppm = Blatt pro Minute) verarbeiten. Dies gibt bereits einen ersten Anhaltspunkt über den Einsatzzweck des Geräts. Allerdings kann man nicht nur die Geschwindigkeit betrachten und den angegebenen Wert mit 60 und der täglichen Stundenanzahl multiplizieren.

Bereits die Belegvorbereitung (Sortieren, Entklammern, Glätten etc.) nimmt Zeit in Anspruch. Es empfiehlt sich immer, einen verhältnismäßig schnellen Scanner zu

wählen, um das Zeitfenster für die Belegerfassung zu verkürzen. Soll zum Beispiel die tägliche Eingangspost von 1.000 Blatt gescannt werden, würde theoretisch ein 25ppm Scanner ausreichen. Allerdings ist dann eine Arbeitskraft etwa zwei Stunden mit dem Einscannen beschäftigt, und mit einem schnelleren Gerät wird nicht nur diese Erfassungszeit gespart, sondern die Belege stehen für die folgenden Arbeiten schneller zur Verfügung.

Außerdem ist zu beachten, dass die angegebene Scangeschwindigkeit in Abhängigkeit vom Papierformat und der gewählten Auflösung variiert, und nicht alle Hersteller gehen bei ihren Geschwindigkeitsangaben von den gleichen Scannereinstellungen aus. Meistens wird der Durchsatz für 200dpi angegeben. Je höher die gewählte Auflösung, desto langsamer wird der Scanner.

Die Qualität der Bilderfassung sollte für unterschiedliche Arten von Dokumenten gewährleistet werden, so dass der Scanner ohne Benutzereingriff stets perfekte Bilder ausgibt. In den Scannern sind unterschiedliche Kameratypen eingebaut, und im Vorfeld der Beschaffung empfiehlt es sich, mit den Originalbelegen zu testen, ob der Dokumentenscanner die geforderten Ergebnisse liefert. Grundsätzlich gilt: je höher die optische Auflösung ist, umso besser die Qualität des entstehenden Bildes. Diese ist nicht mit der Auflösung der Ausgabe zu verwechseln. So kann die Kamera ein Rohbild mit optischen 600dpi scannen, das mit 200dpi ausgegeben werden kann.

Nach wie vor werden die meisten Dokumente in schwarzweiß erfasst und als TIFF Datei abgespeichert. In der Reget reicht eine Ausgabeauflösung von 200dpi aus, um alle wesentlichen Informationen der Belege zu erhalten oder auch elektronisch auszulesen. Die Herausforderung an den Dokumentenscanner besteht darin, ein Farb- oder Graustufenbild ohne Informationsverlust in ein Schwarzweißbild umzuwandeln. Dies geschieht durch „dynamic thresholding“. D.h. ab einer gewissen Schwelle wird ein Bit entweder schwarz oder weiß dargestellt. Durch intelligente Helligkeits- und Kontrastanalyse wird diese Schwelle innerhalb eines Dokuments automatisch verändert, damit können auch bei farbigen oder kontrastschwachen Belegen gute Scannergebnisse erzielt werden. Gerade bei sehr unterschiedlichen Qualitäten der Originalbelege ist diese Schwellenwertfunktion ein Muss. Ebenfalls unerlässlich sind Funktionen wie die automatische Schräglagenkorrektur („deskew“) und „auto cropping“ (Zuschneiden von unterschiedlichen Belegen im Stapel auf die tatsächliche Größe). Die genannten Bildverbesserungen können über den Scannertreiber oder Zusatzprodukte erreicht werden. Der „de facto“-Standard in diesem Segment ist VirtualReScan von Kofax.

Alternativ kann in Farbe gescannt werden und der Markttrend geht auch eindeutig in diese Richtung: Bitonale Geräte sind bis auf einige Ausnahmen im Produktionsbereich nicht mehr verfügbar und alle seit 2006 neu vorgestellten sind farbfähig. Damit kann jederzeit von Schwarzweiß auf Farbscannen umgestellt werden. Die Frage nach Farbe kann beim Scannerkauf mittlerweile als nachrangig betrachtet werden.

Bei älteren oder einfacheren Scannern sollte das Scan-Modul Zusatzfunktionen bereitstellen, die bei modernen Scannern heute per Firmware oder Treiber zur Verfügung stehen: Nämlich dass die Bilder automatisch gerade gerückt werden und dass der Schwarzrand automatisch entfernt wird, Funktionen für die Rauschminderung oder Grauflächenentfernung sind bewusst einzusetzen, da diese ganz leicht i-Punkte

oder Satzzeichen entfernen können. Handelt es sich dabei um Dezimaltrennzeichen auf einer Rechnung, ist sogar der Inhalt des Dokumentes verfälscht. Nur wenige Scanner können auch automatische Rotationen ausführen. Dabei analysiert die Technik die Ausrichtung des aufgedruckten Textes und dreht die Bilder vollkommen automatisch in die richtige Richtung.

Erkennungstechnologien: Barcode, Barcodetypen, Strichcode, OCR, ICR; Abgleich mit vorhandenen Daten

Automatisierte Erkennungstechnologien sollen die Sachbearbeiter vor zu vielen manuellen Eingaben bewahren. Dies kann manuell beim Scannen, durch automatisches Erkennen von Text oder Barcode und durch Hinzufügen fehlender Informationen aus bestehenden DV-Systemen geschehen. Für die automatische Extraktion von Zugriffsinformationen, der sogenannten Indexinformation, sind Techniken wie OCR (Optical Character Recognition) u.ä. erforderlich. Diese können nur unter bestimmten qualitativen Voraussetzungen der Dokumentenvorlagen sicher gewonnen werden.

Zur Verarbeitung von gescannten Belegen werden verschiedene Erkennungstechniken (Recognition - Mustererkennung) eingesetzt. Zu ihnen gehören:

- Barcode: Aufgebrachte Barcodes beim Versenden von Vordrucken können beim Einlesen der Rückläufer automatisiert erkannt und zugeordnet werden.
- OCR (Optical Character Recognition): Hierbei werden die Bildinformationen in maschinenlesbare Zeichen umgesetzt. OCR wird für Maschinenschrift eingesetzt.
- HCR (Handprint Character Recognition). Die Erkennung von Handschriften ist eine Weiterentwicklung von OCR, die jedoch bei Fließtexten immer noch nicht zufriedenstellende Ergebnisse liefert. Beim Auslesen von definierten Feldinhalten ist die Methode doch bereits sehr sicher.
- ICR (Intelligent Character Recognition). ICR ist eine Weiterentwicklung von OCR und HCR, die die Qualität der ausgelesenen Ergebnisse durch Vergleiche, logische Zusammenhänge, Abgleich mit Referenzlisten oder Prüftabellen verbessert.
- OMR (Optical Mark Recognition). liest mit hoher Sicherheit spezielle Markierungen in vordefinierten Feldern aus und hat sich bei Fragenbogenaktionen und anderen Vordrucken bewährt.

Bei der Freiformerkennung müssen die gesuchten Attribute nicht mehr an einer festen Stelle stehen, sondern können irgendwo aufgedruckt sein. Über ein flexibles Layout und ein Dokumentenmodell wird definiert, welche Daten einem bestimmten Dokumententyp zugeordnet sind. Ein flexibles Layout beschreibt die Felder und deren variable Positionen bei einem bestimmten Satz von Dokumententypen. Man erkennt einen Dokumententyp anhand bestimmter Vorgaben und Konventionen. Beispielsweise steht auf einer Rechnung der Rechnungsbetrag, Kontonummer, Rechnungsnummer usw. Der Mensch schaut sich ein ganzes Dokument an und analysiert erst die verschiedenen Elemente und dann einzelne Informationen. Anschließend wird er nach Informationen in der Umgebung der Indexfelder suchen, um zu entscheiden, welche Informationen er in welches Feld der Datenbank eingibt. Ähnlich macht es die Software. Mit Hilfe von Suchelementen, die mit bestimmten Eigenschaften versehen werden und logisch untereinander verknüpft werden, können die gesuchten Elemente gefunden werden. Typische Suchelemente sind Text, Trennlinien, weiße Lücken, Barcodes, Zeichenketten,

Textfragmente, Objektsammlungen, Datum, Telefonnummern, Währungen und sogar komplexe Tabellen. Im Gegensatz zur Erkennung einzelner Elemente, kann durch eine Analyse der Elemente im Kontext die Genauigkeit gesteigert werden. Typische Anwendungen sind die automatische Rechnungseingangserfassung sowie die Klassifikation und Auswertung der Eingangspost. Die Klassifikation ermittelt den Dokumententyp, um die Datenextraktion zu vereinfachen. In einer Personalakte befinden sich beispielsweise ein Bewerbungsschreiben, ein Lebenslauf, Zeugnisse, Beurteilungen und Gehaltsabrechnungen. Für jeden Dokumententyp sind Daten definiert, die das jeweilige Dokument kennzeichnen.

Im Gegensatz zu dem regelbasierten Ansatz sind in den vergangenen Jahren lernende Mustererkennungsverfahren entwickelt worden. Bei den Mustererkennungsverfahren werden der Anwendung viele Musterdokumente vorgeführt, aus denen die Anwendung bestimmte regelmäßige Strukturen entnimmt. Wird ein neues Dokument vorgeführt, wird die Ähnlichkeit zu den gelernten Dokumenten verglichen. Der Einsatz von lernenden Mustererkennungsverfahren wird meist begrenzt auf die Klassifikation und dient als Hilfsmittel der eigentlichen Datenextraktion.

Über automatische Datenbankabfragen lassen sich die erkannten Ergebnisse überprüfen und als „sicher erkannt“ markieren. Auch kann die Datenbank fehlende Indexmerkmale ergänzen. Die Korrekturen lassen sich so auf ein Minimum reduzieren. Nachgelagerte Datenbankabfragen und Konsistenzprüfungen innerhalb der Erfassungslösung stellen zusätzlich sicher, dass die Daten korrekt sind.

Es stellt sich nun die Frage: Welches OCR/ICR-Produkt ist das richtige für die jeweilige Anwendung? Es gibt zahlreiche OCR/ICR-Systeme unterschiedlicher Qualität. Benchmark-Tests und Untersuchungen auf Fehleranfälligkeit führen zu vergleichbaren Ergebnissen, indem man Messungen mit standardisierten Vorlagen unter gleichbleibenden Bedingungen durchführt. Der Aufwand für die Nachbearbeitung der maschinell gelesenen Texte - ein weiteres wichtiges Beurteilungskriterium - kann je nach Anwendung sehr unterschiedlich sein. Von maßgeblicher Bedeutung wird letztlich eine individuell zu erstellende Kosten/Nutzen-Analyse sein.

OCR Beurteilungskriterien

- **Erkennungsmethode**
- **Erkennungsgeschwindigkeit**
- **Schriftgrößen**
- **Erkennung von Nadeldruck (LQ, Draft)**
- **Möglichkeit der Genauigkeitskontrolle**
- **Lexikonunterstützung bei der Erkennung**
- **Integrierte Korrektursoftware**
- **Trainierbarkeit**
- **Automatische Ligaturentrennung**
- **Spaltenerkennung**
- **Hoch- /Querformaterkennung**
- **Komfort der Fensterdefinition**
- **Inputformate (TIFF, PCX etc.)**
- **Outputformate (Textverarbeitung, DTP, Tabellenkalkulation, Datenbank)**
- **Hardwareunterstützung (Co-Prozessor)**
- **Betriebssystem**
- **Benutzeroberfläche**
- **Preis**

Die Eignung der untersuchten Systeme wird anhand von Testmaterial erprobt. Es sollte ein repräsentatives Mustermaterial (Referenzstapel), so wie es später zum Einsatz kommt, zusammengestellt werden. Alle Systeme werden der gleichen Testprozedur unterworfen.

Besondere Erwähnung in diesem Zusammenhang muss der Einsatz von RFID (Radio Frequency Identity) erwähnt werden, der sicherlich in naher Zukunft verstärkt im Umfeld von ECM-Anwendungen zu finden sein wird. Informationsobjekte im ECM-Umfeld können digitaler oder körperlicher Natur sein. Einige in körperlicher Form vorhandene Informationsobjekte, wie z.B. Dokumente in Papierform, lassen sich digitalisieren. Bei einer Reihe von Objekten, z.B. Blutproben, Gemälden, Postpaketen, Bauteilen etc. ist diese Digitalisierung nicht möglich. In der Regel wird bei diesen rein körperlich vorliegenden Objekten nur der Standort im ECM abgespeichert.

Der Einsatz von RFID erfordert zwei Komponenten: den RFID-Tag als Prozessor, Speicher, Sende- und Empfangseinrichtung und den RFID-Reader. Der Reader erzeugt ein elektromagnetisches Feld, welches die Antenne des Transponders empfängt und es als Befehl an den RFID-Tag weiterleitet. Der Transponder sendet Antwort an das elektromagnetische Feld und der Reader interpretiert die Antwort als Antwortdaten.

Denkbare Anwendungsformen von RFID im ECM-Kontext könnten sein:

- Verfolgung und Archivierung von Proben aller Art
- Verfolgung und Archivierung von Urkunden, die im Original vorliegen müssen
- Verfolgen und Archivierung von Büchern und Zeitschriften
- Bessere Unterstützung des Posteingangs
- Leichteres Finden und Zuordnen verloren gegangener Objekte
- Compliance Unterstützung durch automatischen Vergleich von Dokumentation und tatsächlich umgesetzter Realität (wurden die geplanten Teile eingebaut und befindet sich keine Fälschung darunter z.B. Schiffbau, Anlagenbau, Flugzeugbau etc.)
- Auslösen von Alarmen, bei vorher definierten Ereignissen (z.B. zu hohe Luftfeuchtigkeit)
- Zeitgewinn bei der Beseitigung von Schäden (z.B. bei Temperaturüberschreitung kann ohne visuelle Überprüfung festgestellt werden, welche Objekte in einem bestimmten Stapel, Raum oder anderen Einheit betroffen sind)

Formulare, Formularmanagement, Formularverarbeitung; Design von Formularen

Eine besondere Bedeutung bei der Erfassung von Merkmalen für die Charakterisierung von DMS-Objekten spielen Formulare oder Vordrucke. Ein Formular ist ein standardisiertes Mittel zur Erfassung, Ansicht und Aufbereitung von Daten. Formulare sind Vervielfältigungen, die durch Eintragungen zu ergänzen sind und der Bearbeitung häufig auftretender, gleichartiger Geschäftsfälle dienen.

Bei der Erfassung von Formularen werden heute noch zwei Gruppen von Techniken unterschieden, obwohl der Informationsinhalt und der Charakter der Dokumente gleich sein kann:

- Forms Processing (Vordruckverarbeitung). Das „Forms Processing“ bezeichnet die Erfassung von industriell oder individuell gedruckten Vordrucken mittels Scannen. Hierbei kommen anschließend häufig die vorne beschriebenen Erkennungstechniken zum Einsatz, da gut gestaltete Vordrucke eine weitgehend automatische Verarbeitung ermöglichen.
- E-Forms / Web-Forms (Verarbeitung elektronischer Formulare). Bei der Erfassung elektronischer Formulare ist eine automatische Verarbeitung möglich, wenn Layout, Struktur, Logik und Inhalte dem Erfassungssystem bekannt sind.

Vordrucke müssen immer einheitlich aussehen und gleichartig verarbeitet werden können. Dabei ist es gleichgültig, ob sie industriell vorgedruckt (herkömmliches Papierformular), individuell vorgedruckt (Ausdruck auf Laserdrucker) oder am Bildschirm angezeigt und mit Daten ausgedruckt (im LAN mit der Anwendung des Sachbearbeiters, als interaktives PDF oder im Internet als Web-Formular) dargestellt werden. Ein Vordruck gleichen Inhalts und Rechtscharakters kann in unterschiedlicher Form vorliegen und muss aber gleich behandelt werden.

Formulare haben dann besondere Vorteile, wenn es sich um rücklaufende Eingangspost handelt. Denn diese ist dann so strukturiert, wie die Organisation, die sie verarbeiten will, benötigt. Formulare und Vordrucke bleiben weiterhin eines der wichtigsten Mittel der Informationserhebung, Informationsorganisation und Prozesssteuerung. Inzwischen geht es nicht mehr nur um die Verarbeitung von Papiervordrucken und deren Datenextraktion. Ein Schwerpunkt ist die Identifikation von elektronischen Formularen, PDF-Formularen und Vordrucken in Papierform um durchgängige Prozesse umsetzen zu können. Dabei kommt eigenständigen Textbaustein- und Formularmanagement-Lösungen mit entsprechender Versionierung, Synchronisation mit Datenmodellen und weiterer Verwaltungsfunktionalität eine wichtige Bedeutung zu.

Übernahme von Daten und Dateien: COLD, Listenformate

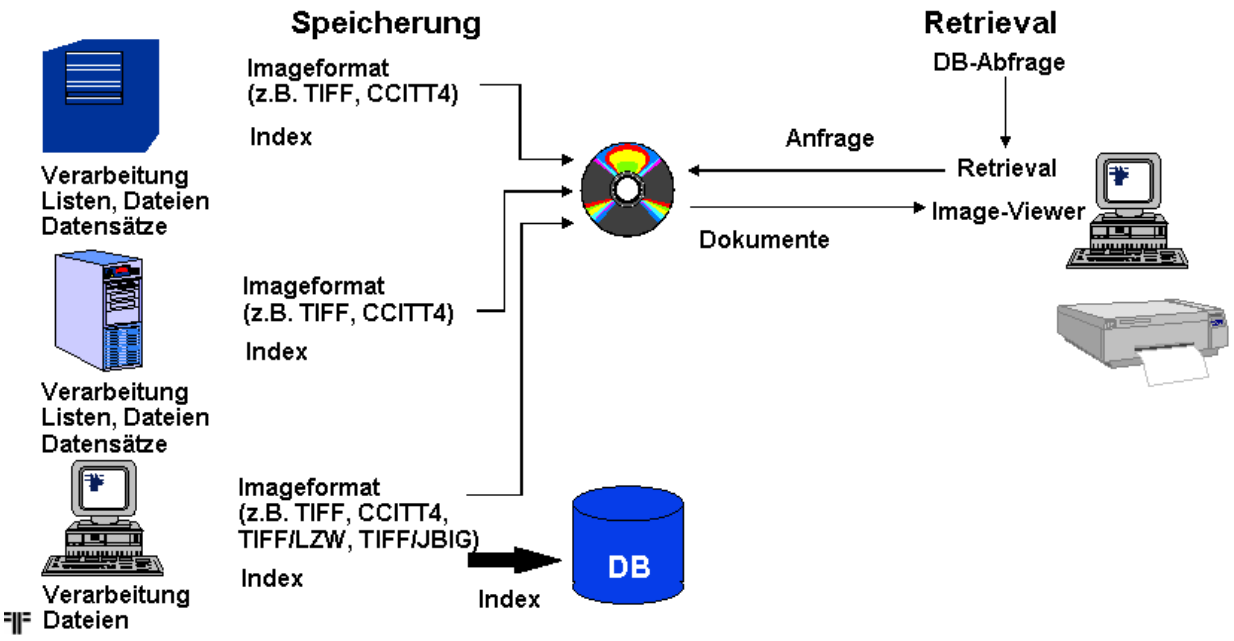
Wie bereits erwähnt sind Belege, die selbst erzeugt werden und deren Aufbau und Datenstruktur damit bekannt sind, besonders für eine automatische Zuführung in ein ECM-System geeignet. Das Verfahren mit dem dies geschieht, nennt man COLD. COLD/ERM sind Verfahren zur automatisierten Verarbeitung von strukturierten Eingangsdateien. Der Begriff COLD steht ursprünglich für Computer Output on LaserDisk und hat sich gehalten, obwohl das Medium LaserDisk seit Jahren nicht mehr am Markt ist. Das Akronym ERM steht für Enterprise Report Management. In beiden Fällen geht es darum, angelieferte Ausgabedateien auf Basis vorhandener Strukturinformationen so aufzubereiten, dass sie unabhängig vom erzeugenden System indiziert und an eine Speicherkomponente wie eine dynamische Ablage (Store) oder ein Archiv (Preserve) übergeben werden können. Die „Aggregation“ stellt einen Kombinationsprozess von Dateneingaben verschiedener Erstellungs-, Erfassungs- und zuliefernden Anwendungen dar. Zweck ist die Zusammenführung und Vereinheitlichung von Informationen aus unterschiedlichen Quellen, um sie strukturiert und einheitlich formatiert an die Speicher- und Bearbeitungssysteme zu übergeben.

Ausgangsdokumente werden üblicherweise auf Druckern erzeugt. Deswegen bietet es sich an, die Druckdaten selbst zu archivieren. Da diese Daten typischerweise in größeren Läufen erstellt werden (Spool-Dateien), hat sich das COLD-Verfahren für die

Archivierung dieser Druckdaten entwickelt. Dabei kommen unterschiedliche Strategien zum Einsatz: Abspeicherung als Rohdaten, die durch Aufruf von Layout-Ressourcen in ein anzeigefähiges Dokument gewandelt werden, Abspeicherung in Listenform für strukturierte Berichte und Auswertungen sowie strukturierte Anzeige der Daten in Anwendungen und Wandlung des Datenstroms in einzelne PDF- oder TIFF-Objekte, die den versendeten Ausgangsbriefen bildlich entsprechen.
Seit einiger Zeit sind die Anforderungen auch im Massendruck immer komplexer. Entsprechend wurden optimierte Druckdatenströme wie PCL, Postscript, BETA93 oder AFP entwickelt, die aufwändige Layouts ermöglichen. Die Speicherung solcher Druckdaten als einzelne Datei ist daher nicht mehr in jedem Fall praktikabel. Entsprechend bieten moderne COLD-Systeme folgende Grundfunktionen: Separieren der Dokumente mit unterschiedlicher Seitenzahl, Index-Ermittlung über frei definierbare Logik, Konvertierung der Druckdaten in ein archivgeeignetes Format wie zum Beispiel TIFF oder PDF/A und Erstellen von speziellen Importdateien für unterschiedliche Archivsysteme.

AFP ist ein von IBM entwickeltes Format für den Druckdatenstrom im Rahmen der Herstellung von Massendruckstücken. Ein AFP-Datenstrom besteht aus Verbunddokumenten (MO:DCA mixed object document content architecture) mit Text, grafischem Inhalt, Schriften und Barcodes. Für die Bildschirmanzeige ist ein AFP-Viewer erforderlich. BETA93 ist ein Output-Verwaltungssystem, mit dem Listen (Ergebnisunterlagen) online zur Verfügung gestellt werden anstatt auf Papier. Es verwaltet, verteilt und archiviert Listen automatisch. PCL (von Hewlett-Packard entwickelt) ist eine Befehlssprache zum Steuern von Laserdruckern und liegt in seiner Komplexität zwischen ASCII (welche nur die einfachsten Kommandos erlaubt, wie z.B. Zeilenvorlauf) und PostScript (welches eine eigene, komplexe Programmiersprache ist und einen Interpreter voraussetzt).

Archivierung strukturierter und unstrukturierter Daten im Imageformat



Die COLD-Archivierung stellt eine sehr wichtige Komponente von Enterprise-Content-Management-Systemen dar. COLD dient hier sowohl für die Aufbereitung der Ausgabe für verschiedene Formate als auch als Eingang für die Bereitstellung der Postausgänge in einer gesamtheitlichen Sicht auf alle elektronischen Dokumente (virtuelle Akte). COLD-Verfahren werden ferner für Migration von Datenspeichern und die automatisierte Überführung von Bewegungsdaten in Dokumentenmanagement-Lösungen benutzt. Ebenso werden auf diesem Weg automatisch MS Office-Dokumente abgelegt.

Eine besondere Erwähnung in diesem Zusammenhang soll hier die Archivierung von SAP-Daten finden. Auf Grund der sehr großen Verbreitung als ERP-System hat dieser Aspekt eine herausragende Bedeutung. SAP verfügt zwar selbst auch über mehr oder weniger umfangreiche ECM-Funktionen, dennoch werden in entsprechenden ECM-Umgebungen SAP-Belege (Rechnungen, Lieferscheine, Buchungsbelege etc.) im COLD-Verfahren archiviert.

Im Zusammenhang mit COLD kann es nicht ausbleiben, das Thema EDIFACT zu streifen. EDIFACT ist die Abkürzung für Electronic Data Interchange For Administration, Commerce and Transport. EDIFACT ist ein branchenübergreifender internationaler Standard für das Format elektronischer Daten im Geschäftsverkehr. Die bei uns gebräuchlichen EDIFACT-Sätze bestehen aus einem Umschlag, den man sich als ein Briefkuvert vorstellen kann. In diesem Umschlag stehen jeweils vereinbarte Codenummern für Absender und Empfänger, sowie Nachrichteninhalt, Zeiten zur Rückverfolgung, sowie Prüfelemente. Eine Nachricht selbst besteht aus Segmenten, Datenelementgruppen und Datenelementen. EDIFACT ist ein Standard für das Datenformat, nicht für die Übertragung der Daten, das heißt im Prinzip können EDIFACT-Nachrichten über jedes Medium/Protokoll ausgetauscht werden, das zur Übertragung elektronischer Daten benutzt werden kann. Ursprünglich wurde EDIFACT auf Standleitungen eingesetzt. Es gab auch erfolgreiche Projekte, die EDIFACT-Nachrichten per Diskette oder Magnetband transportierten. Auch das Internet kann natürlich für EDIFACT genutzt werden.
EDIFACT-Nachrichten sind natürlich prädestiniert für die Archivierung per COLD. Es müssen bestimmte Voraussetzungen geschaffen werden. Entweder sind die beteiligten Anwendungen in der Lage, EDIFACT-Nachrichten zu erzeugen oder zu verarbeiten, oder es wird ein Konverter dazwischengeschaltet, der die Daten entsprechend umwandelt. Ein moderner Konverter kann heute jedes Format in jedes Format umwandeln. Zusätzlich wird dann eine Steuerung verwendet, die den Kommunikationsprozess von der Partnerverwaltung, der Tabellenverwaltung, dem Logging und der Archivierung vollautomatisch übernimmt. EDIFACT ist ein Format, das die ganz überwiegende Mehrheit aller Geschäftspapiere beschreibt. Es ist notwendig, zwischen den Partnern genaue Vereinbarungen über Dateninhalte zu treffen, die die Kannfelder und Mussfelder in ausgewählten Segmenten festlegt.

Formate JPEG 2000, AFP, PDF/A, TIFF, XML etc.; Format-Konvertierung

Die Speicherung der zu verwaltenden Objekte innerhalb eines ECM-Systems ist ein sehr bedeutender Komplex, da das Format letztendlich die Verwendbarkeit und Darstellung des betreffenden Objekts bestimmt. Da es eine Vielzahl verschiedener Formate gibt und mit Dokumenten nur dann umgehen kann, wenn man z.B. einen

geeigneten Viewer besitzt, ist der Ruf nach standardisierten Formaten nur zu verstehen. Die Dokumentenformate lassen sich in folgende Gruppen zusammenfassen:

- Rasterformate – sind alle Formate, in denen der darzustellende Inhalt mit einzelnen Bildpunkten unterschiedlicher Farbe wiedergegeben wird. Beispiele sind TIFF und JPEG.
- Auszeichnungsformate – sind alle Formate, in denen das Layout der jeweiligen Objekte mittels lesbarer Kennzeichnungen (Tags) festgelegt wird. Beispiele sind XML und HTML.
- Mixed Object-Formate – beschreiben jeden Objekttyp (Text, Grafik, Barcode etc.) in einer eigenen, seine Besonderheiten berücksichtigenden Weise. Beispiele sind, PDF, PDF/A, PCL und AFP.

Die im Umfeld von ECM-Lösungen gebräuchlichsten Formate sind wohl TIFF, JPEG, XML und PDF bzw. PDF/A.

TIFF wurde als Format zur Speicherung von Bilddaten von Aldus (später: Adobe) und Microsoft für gescannte Rastergrafiken für die Farbseparation entwickelt. Es unterstützt S/W-Bilder und Bilder mit Grauwerten oder Farben sowie einseitige (single-page-TIFF) und mehrseitige Dokumente (multi-page-TIFF). Die Komprimierung von S/W-Bildern kann verlustfrei erfolgen durch z.B. Fax Gruppe 3 oder 4. Farbbilder werden verlustfrei komprimiert durch LZW-Verfahren oder verlustbehaftet durch JPEG-Verfahren. TIFF kann mit Standard-Viewern angezeigt werden. Manchmal gibt es dabei durch verschieden Versionen Kompatibilitätsprobleme. Es ist das bisher am meisten verwendete Speicherformat in DMS/ECM-Anwendungen.
JPEG oder besser: JPEG2000 ist seit 2001 offizieller weltweiter Kompressionsstandard der ISO für statische Bilder. Er basiert auf einer Wavelet-Kompression und ist geeignet für die elektronische Speicherung und Archivierung sowie die Langzeitarchivierung.
XML ist ein Sprachenstandard zur Strukturierung und Beschreibung von Daten. Die Speicherung geschieht nicht im Binärformat, sondern im Textformat. Es stellt eine echte Untermenge von SGML.
PDF wurde von Adobe entwickelt als Dateiformat mit dem es möglich ist, elektronische Dokumente unabhängig von Textverarbeitungsprogramm und/oder Betriebssystem originalgetreu zu nutzen. Es ist ein statisches Format und beinhaltet alle Layout- und Schriftinformationen des Originals. Es unterstützt gleichzeitig eine flexible Architektur für digitale Unterschriften. Das PDF-Format ist eine Weiterentwicklung von Postscript und wird mit dem kostenlosen Acrobat Reader oder anderen PDF-Viewern gelesen.
PDF/A hat eine besondere Bedeutung, da sich hier ein Standard entwickelt hat, den alle Experten als das Format für die Langzeitarchivierung einstufen, das überall verwendet werden kann. Betrieben wurde die Entwicklung von den US Branchenverbänden NPES und AIIM. Erste Entwürfe zum Standard ISO 19005-1 liegen vor.
Als Ziele für die Langzeitarchivierung sind sowohl Versionsunabhängigkeit, Revisionssicherheit und Unabhängigkeit von Adobe definiert. Technisch gesehen handelt es sich hier um eine erweiterte und auch beschränkte PDF1.4-Definition. Es sind z. B. eingebettete Scripte nicht erlaubt. Der Acrobat-Reader ist verfügbar aber noch kein PDF/A Viewer. Wesentliche Anforderungen, die an dieses Format gestellt wurden sind:

- Geräte-, Software- und Versionsunabhängigkeit, so dass die Inhalte immer gleich dargestellt werden
- „Self Contained“, beinhaltet alle Komponenten, die zur Darstellung nötig sind, in der Datei inklusive Fonts
- „Self Documented“, die Dateien beschreiben sich inhaltlich und dokumentieren sich selbst über wiederum standardisierte Metadaten
- Transparenz, eine PDF/A-kompatible Datei ist mit einfachen Mitteln analysierbar

Es gibt bereits viele Organisationen und Unternehmen, die PDF/A als generelles Speicherformat einsetzen. Beispielhaft sei hier die DAK Krankenkasse genannt.

Sehr viel detaillierte Informationen gibt es hierzu beim „PDF/A Competence Center“ unter www.pdfa.org.

Vor-/Nachteile verbreiteter Formate

Merkmale	B	TIFF/G3 oder G4	B	PDF	B	Prop. Formate
Volltextfähig	-	Nein, erst durch OCR/ICR	+	Ja (aber nicht für TIFF-in-PDF)	+	Ja, wenn Filter verfügbar
Ressourcen-Probleme	+	Nein	Ø	Manchmal	Ø	Manchmal
Verfügbarkeit von Viewern	Ø	Unterschiedliche Viewer, nicht alle Plattformen	+	Sehr gut	-	Nur für den Anwender der Autorensoftware oder durch 3rd-Party Formatviewer
Veränderbarkeit	Ø	Einfach durch Editoren, System muss schützen	Ø	Einfach durch Editoren, System muss schützen	Ø	Einfach durch Editoren, System muss schützen
Zukunft-sicherheit des Formates	+	Hoch	+	Hoch	Ø	Hoch bei Einfachformaten (ASCII) und weit verbreiteten Formaten (Bsp. Word). Riskant bei anderen
Kann farbige Texte enthalten	-	G3/G4 nur bitional. TIFF mit JPEG möglich	+	Ja	+	Ja
Kann farbige Grafiken enthalten	-	G3/G4 nur bitional. TIFF mit JPEG möglich	+	Ja (positiv: Embedded Bitmaps werden komprimiert	+	Ja (Achtung: Kompression!)

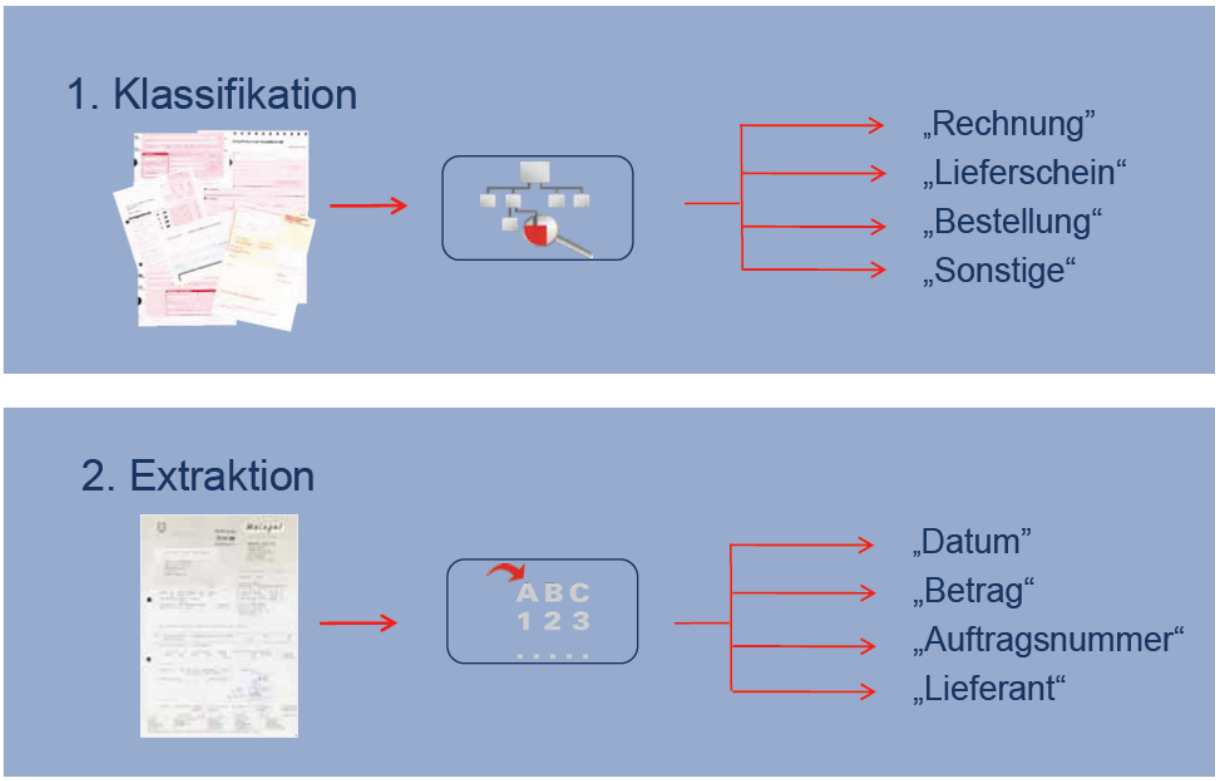
Nicht immer stehen Objekte in dem Format zur Verfügung, welches in der aktuellen Anwendung verwendbar ist. Deswegen müssen in diesen Situationen Konvertierungen vorgenommen werden. Es gibt verschiedenste Zeitpunkte und Gründe für die Anforderung, Dokumente in ein anderes Format zu konvertieren. So kann es vorkommen, dass für eine Anzeige am Bildschirm das Dokument konvertiert werden muss. Dies geschieht unmittelbar während des Viewing-Prozesses. Das Original bleibt dabei unverändert. Vielleicht muss eine Konvertierung für eine Langzeitarchivierung vorgenommen werden. Hierbei wird das Originalformat verändert. In manchen Fällen müssen sogar mehrere Formate des gleichen Inhalts unter gleichem Index archiviert werden. Dies kann der Fall sein, wenn S/W- und Farb-Dokumente abgelegt werden

müssen. Um im Web Dokumente verwenden zu können, ist es erforderlich, diese zu konvertieren.

Ein häufig verwendetes Format ist hier Adobe Flash als eine proprietäre integrierte Entwicklungsumgebung zur Erstellung multimedialer Inhalte, so genannter „Flash-Filme“. Die resultierenden Dateien liegen im SWF-Format vor, einem auf Vektorgrafiken basierenden Grafik- und Animationsformat. Das Kürzel SWF steht dabei für Shockwave Flash. Um Flash-Dateien betrachten zu können, ist das proprietäre Abspielprogramm Flash Player erforderlich, das auch als Webbrowserplugin eingebunden werden kann. Flash findet heutzutage auf vielen Webseiten Gebrauch z.B. als Werbebanner.

Automatische Klassifikation, Taxonomien

Generell versteht man unter Klassifikation die Bildung von Dokumentenklassen oder Informationsobjektklassen. Diese dient zur Gruppierung von Objekten mit gleichen Attributen oder Eigenschaften. Die Nutzung von Dokumentenklassen ist eine der wesentlichen Eigenschaften von Systemen um Dokumente und Informationsobjekte zu schützen, zu strukturieren und zu ordnen, in geeigneter Form in elektronischen Akten zu visualisieren und effizient zu verwalten. Typische Attribute von Dokumenten- oder Informationsobjektklassen sind Schlagworte, Ordnungskriterien, Berechtigungen, Speicherorte, Aufbewahrungsfristen, Vernichtungszeitpunkte. Den Prozess der Zuordnung von Objekten zu einer Klasse bezeichnet man als Klassifizierung oder auch Indizierung oder Attributierung.



Wie bereits weiter oben ausgeführt, ist es ein verständliches Anliegen der Anwender, diesen Prozess möglichst zu automatisieren und eingehende, elektronische und

papierbasierte Dokumente automatisch zu indizieren und zuzuordnen. Hierzu werden nach der Erfassung durch den Scanner die Faksimiles mit OCR/ICR-Techniken interpretiert und anschließend werden die Indexmerkmale nach vordefinierten Schemata herausgefiltert, geprüft und mit Stammdaten abgeglichen. Selbstlernende Programme erlauben die Generierung von Strukturen, Aufbau von Ordnungssystematiken und Zuordnungen anhand der Dokumentinformationen. Elektronische Informationsobjekte (z.B. per OCR-gewandelte Faksimiles, Office-Dateien oder Ausgabedateien) enthalten Informationen. Auf Basis dieser Informationen können Programme zur automatischen Klassifikation selbstständig Index-, Zuordnungs- und Weiterleitungsdaten extrahieren. Solche Systeme können auf Basis vordefinierter Kriterien oder selbstlernend Informationen auswerten. Zur Anwendung können hierbei neuronale Netze oder regelbasierte Ansätze kommen. Neuronale Netze sind selbstlernende Mechanismen bzw. künstliche Intelligenz. Auf Grund einer definierten Lernmenge können alle weiteren Informationen automatisch zugeordnet werden. Der Effekt des Übertrainierens kann diese Systeme allerdings in unkontrollierbare Zustände versetzen.
Regelbasierte Ansätze sind alle Merkmale, die einen bestimmten Dokumententyp beschreiben, müssen durch eine Regel beschrieben werden. Bei sehr heterogenen Dokumenten können erhebliche Administrationsaufwände für die Erstellung der Regeln entstehen.

Automatisches Lernen
- Sinnvoll bei vielen verschiedenen Dokumententypen
- Reduziert den Erstaufwand beim Setup
- Führt zu untransparentem Verhalten
- Schwer zu optimieren und zu pflegen
- Funktioniert bei der Informationsextraktion nur eingeschränkt

Regelbasiertes Lernen
- Sinnvoll bei geringer oder mittlerer Anzahl von Dokumententypen
- Dokumentanalyse durch qualifiziertes Personal
- Einfach zu verstehen und zu optimieren
- Kann auch spezielle Sonderfälle behandeln
- Funktioniert sowohl bei der Klassifikation als auch bei der Informationsextraktion

Es ist verständlich, dass solche Automatismen nicht immer fehlerfrei ablaufen. Es können Erkennungsfehler auf Zeichenebene, Feldebene, Dokumentebene vorkommen. Die Fehlerrate nimmt mit jeder Ebene exponentiell zu. Als Fehler oder Probleme kann es beispielsweise geben:

- Nicht gefunden: das Feld steht auf dem Dokument wurde aber nicht erkannt
- Falsch gefunden: das Lieferdatum wurde als Rechnungsdatum interpretiert
- Falsch gelesen: 12.08.2004 wurde als 12.03.2004 erkannt
- Fehlerhafterweise gefunden: auf dem Dokument fehlt die Information, trotzdem wird aber ein Wert geliefert.
- Normierung beachten: Sind 12.08.2004 und 12-8-04 identisch?
- Andere Beispiele: Goethestr. → Goethe Straße oder ++49 7531 874259 → 7531 874259
- Felder haben unterschiedliches Gewicht: Adresse hat mehr Zeichen als ein Betrag / besteht ein Datum aus einem oder aus drei Feldern?

In einer automatischen Nachverarbeitung kann unter Zuhilfenahme von Kontextwissen eine Korrektur von Fehlern vorgenommen werden. In einem geometrischen Kontext werden Beträge besonders behandelt oder es wird ein Liniensystem anhand der Erkennung gebildet. Dieses erlaubt es solche Erkennungsunterscheidungen wie z.B. bei „9“ oder „G“, „,“ oder „l“, „o“ oder „O“ vorzunehmen. Im logischen Kontext lassen sich Groß- und Kleinbuchstabenworte eindeutig interpretieren. Beispielsweise wird „H A L L [0 | O | o]“ eindeutig als „HALLO“ erkannt. Als lexikalischer Kontext wird der Abgleich gegen ein Wörterbuch bezeichnet. Dies nennt man auch taxonomischer Ansatz.

Die Taxonomie in Bezug auf Dokumente steht für ein Klassifikationssystem, eine Systematik oder den Vorgang des Klassifizierens zur Schaffung von Ordnungs- und Ablagestrukturen mit Beziehungen zwischen Klassen und den zugeordneten Objekten. In diesem Umfeld ist auch der Begriff Ontologie angesiedelt. Ontologien dienen der Bildung von Kategorien, denen Einzelobjekte zugeordnet werden. Zusammengehalten wird eine Ontologie durch Interferenz- und Integritätsregeln, die die Struktur, die Beziehungen und die Regeln selbst in einem geschlossenen kausalen System logisch und nachvollziehbar definieren. Der Unterschied zwischen Taxonomie und Ontologie ist, dass die Ontologie ein Netzwerk von Informationen mit logischen Relationen darstellt, während die Taxonomie eine hierarchische Untergliederung bildet. In modernen Systemen kann man aber auch eine Taxonomie als hierarchische Abbildung einer auf Relationen aufgebauten Ontologie betrachten, die Restriktionen klassischer hierarchischer Ordnungsprinzipien, in denen sich ein Begriff oder ein Objekt immer nur an einer Position in einer Hierarchie befinden kann, gilt in IT-Systemen nicht mehr. Ähnliches gilt übrigens auch für hierarchisch aufgebaute Thesauri, wo mittels Crosslinks nach ISO 2788 die rein hierarchischen Zuordnungen „durchbrochen“ werden können. Die Nomenklatur ist eine Anwendung der Taxonomie.

Ein kontrollierter und/oder prüfbarer Wortschatz, der als Auswahllisten und hierarchische Thesauri bereitgestellt wird, sichert die Einheitlichkeit der Indizierung. Im Gegensatz zu herkömmlichen Ablageorganisationen kann jedes Dokument mit mehreren Begriffen verbunden werden, was später die Suche erheblich vereinfacht.

Besondere Szenarien und Anwendungsbeispiele

Die bisherigen Ausführungen zum Thema „Capture“ werden abgeschlossen mit einigen kurzen Szenarien und Anwendungsbeispielen, wie sie uns in der Praxis begegnen.
.

Posteingangslösungen

Die mit am Häufigsten anzutreffende Anwendung in ECM-Lösungen ist die Verarbeitung des Posteingangs. Sie ist auch eine der Anwendungen, die auf Grund der Verschiedenartigkeit der Belege mit die höchsten Anforderungen an das Capturing stellt. Automatisierte Posteingangserfassung mit maschineller Indizierung dient zur Überwindung des Flaschenhalses der Erfassung. Bei diesen Anwendungen ist immer der Umfang der Automatisierbarkeit zu prüfen. Organisation besonders bei stufenweiser Einführung, Zeitpunkt der Bereitstellung, Abhängigkeit von der Verfügbarkeit und Prozentsatz „richtig“ erkannter Daten und Aufwand für Korrekturen sind Punkte, denen besondere Aufmerksamkeit geschenkt werden muss. Wo es möglich ist, sind alle Einflussmöglichkeiten auf selbst erstelltes Schriftgut (Vordrucke, Formulare,

Individualbriefe) auszunutzen. Letztendlich sind alle rechtlichen Grundlagen wie Unterschriften, Papiervernichtung und Elektronisches Posteingangsbuch (zertifizierte Zeitstempel) sorgsam zu berücksichtigen. Ein „einheitlicher Postkorb“ ist das, was die Anwender eigentlich wollen. Das bedeutet:

- Alle Nachrichten aus den Quellen Posteingang, Vorgangsbearbeitung, interne E-Mail, Internet-Mail, Fax, Datenbankrecherchen, Sprachaufzeichnung etc. in nur einem Posteingangsordner
- Der Medienbruch zwischen derzeit verschiedenen Ordnern („man muss wissen, wo die Information ist“) soll vermieden werden.
- Eine einheitliche, strukturierte Benutzeroberfläche mit Dokumentenmanagement-Funktionalität soll angeboten werden.

Rechnungseingangsverarbeitung
Eine Anwendung mit sehr viel Einsparungspotenzial ist die Verarbeitung von eingehenden Rechnungen. Hier werden nach dem Scannen die Rechnungen einer Dokumentenanalyse unterzogen, diverse Daten extrahiert und abschließend in Verbindung mit dem bestehenden ERP-System die Rechnungen eingebucht. Als Daten werden extrahiert:

- Kopfdaten (Identifizieren des Buchungsvorgangs, Lieferant, Rechnungsdatum, Währung etc.)
- Positionsdaten (Artikelangaben, Menge, Einzelpreis, exakte Kontierungsangaben, Steuerangaben, Beschreibung, Zahlungsbedingungen, Skonto etc.)
- Zusatzangaben (Kostenstelle, Auftragsdaten, Projektdaten, Aufteilungsschlüssel etc.)

Schwierigkeiten, die eine 100%ige automatische Verbuchung behindern tauchen dann auf, wenn Rechnungen von noch nicht bekannten Kunden eingehen, noch keine Kunden-Nummer angelegt ist, das Rechnungs-Layouts sich geändert hat oder große Sammelrechnungen geteilt werden müssen.

Scannen mit elektronischer Signatur
Generell werden elektronische Rechnungen durch die elektronische Signatur effizienter. Denn nur dadurch ist gewährleistet, dass die Vorsteuer abgezogen werden darf (UStG §14), dass umfangreicher Papieraustausch (teilweise zusätzlich zur elektronischen Rechnung) und Aufwendungen für Papierhandling (Portokosten) vermieden werden. Es sind Lösungen sowohl für Ein- als auch für Ausgangsrechnungen verfügbar.
Für die Nutzung der elektronischen Signatur beim Scannen wird verlangt, dass die korrekte Verarbeitung, Indizierung und Qualitätskontrolle jedes Einzeldokuments durch das Scan-Personal mit der persönlichen Signaturkarte bestätigt wird. Welchen Wert hat diese Signatur? Sie gilt nur als Bestätigung, dass vollständig und lesbar erfasst wurde. Es gibt keine Beziehung zum Absender bzw. Nutzer des Dokuments. Diese Vorgehensweise ist bei Organisationen, die der Sozialgesetzgebung unterliegen bindend.
Die elektronische Signatur verbreitet sich zunehmend, die Signatur findet zunehmend Akzeptanz, besonders bei elektronischen Rechnungen. Es besteht bereits ein großes Angebot an Produkten. Die Öffentliche Verwaltung ist Vorreiter bei Fachanwendungen.

E-Mail-Management

E-Mail-Management rückt immer stärker in den Fokus der Anwender. Immer mehr eigenständige Produkte kommen auf den Markt. Noch beherrschen reine E-Mail-Archivierungslösungen den Markt, stellen jedoch eine Sackgasse dar. Langsam wird den Anwendern bewusst, dass sie weitere Insellösungen schaffen und eine Integration in vorhandene Ablagen in Kunden-, Vorgangs- oder Sachakten benötigen. Die Bürokommunikationsprogramme quellen über und werden gleichzeitig immer größer und komplexer. Der Spam-Anteil bzw. der Anteil an unnötigen Mails und der Aufwand für die Systemadministration steigen. Bei E-Mails ist die Identifikation der aufbewahrungspflichtigen, der aufbewahrungswürdigen und der übrigen Dokumente besonders aufwändig, insbesondere wenn private Nutzung erlaubt ist. Häufig werden E-Mails und/oder Attachements gedruckt und dann abgelegt.
Verschiedene Alternativen werden bei Anwendern eingesetzt. Die vollständige Archivierung liefert mit Sicherheit Konformität mit den rechtlichen Erfordernissen, bietet die Möglichkeit der Prozessintegration und hat kaum Einfluss auf die Arbeitsweise der Nutzer. Sie ist aber zugleich umfangreichste, aufwändigste und damit meist kostenintensivste Lösung. Die vollständige Archivierung mit Separierung der Anlagen führt zu wesentlicher Reduzierung der Mailboxgrößen, erhöht damit die Performance und benötigt weniger Hardware. Bei der selektiven Archivierung werden nicht immer alle rechtlichen Anforderungen erfüllt. Nach einem Crash des Mailsystems ist eine Zuordnung der archivierten Anhänge zu Mails meist schwierig. Ein Offline-Rückgriff auf die Anhänge ist nicht möglich. In der Realität finden sich natürlich auch Kombinationen der Verfahren. Alle diese Verfahren können anwender- oder system-getrieben eingesetzt werden.

Fax als „Scannen an entferntem Ort“

Ein Fax-Gerät ist eigentlich nichts anderes als ein Scanner. An einem entfernten Ort werden Dokumente in ein Faksimile-Format gebracht und an einen Empfänger verschickt. Da diese Dokumente bereits in digitaler Form eingehen, können sie bei entsprechender Infrastruktur vollautomatisch erfasst werden. Dies führt zu erheblichen Einsparungen durch weniger Kosten für Ausdrucke, Wegfall der Zeit beim Scannen, Reduzierung der Zeit für Indizierung und Arbeitsvorbereitung. Als zusätzlicher Nutzen ergibt sich kein Medienbruch, kein Zeitverlust bei der Weiterverarbeitung der Schriftstücke, Prozessoptimierung und Nutzung aktueller technischer Möglichkeiten des ECM-Systems.

Modulares Scan-Subsystem

Hierunter ist ein Subsystem zu verstehen, dass auf einem Bus einzelne Module, die über Profile gesteuert, zugeschaltet oder nicht genutzt werden, den vollständigen Erfassungsprozess bis zur Übergabe an Folgesysteme erledigt. Bei diesem System-Aufbau legt der Administrator zentral fest, welche Arbeiten an den entfernten Standorten erledigt werden sollen. Die Dokumente können sofort nach dem Scannen übermittelt werden oder sie können in der entfernten Niederlassung weiterverarbeitet werden. Mitarbeiter vor Ort, die mit den Dokumenten vertraut sind, übernehmen die Indexierung und Validierung.
Vorteile dieses konzernweiten Schriftgutmanagements sind z.B. in einem echten Einsatz eine einheitliche Plattform für die unterschiedlichen Anforderungen der Tochterunternehmen, ausreichende Flexibilität trotz einheitlicher Plattform, um eine

Vielfalt von Anforderungen abdecken zu können, Minimierung des Wartungsaufwandes und der Ausfallzeiten sowie die schnelle Reaktion auf neue Anforderungen.

Zentrales vs. dezentrales Scannen

Generell lassen sich beim Scannen zwei Strategien unterscheiden. Bei der zentralen Erfassung gibt es eine Stelle im Unternehmen, wo die Post erfasst wird (Scan- oder Post-Stelle). Bei der dezentralen Erfassung gibt es einige bis zu einer Vielzahl von Scannern, die über das Unternehmen verteilt oder sogar in ausgelagerten Stellen (z.B. Baustellen, Filialen) installiert sind. Die entsprechenden Sachbearbeiter erfassen die Belege vor Ort.
Zentrale Posteingangslösungen machen bei höherem Papieraufkommen Sinn. Die technische Auslegung der Lösung einschließlich redundanter Komponenten ist an einer Stelle konzentriert. Eine höhere Qualität bei der Erfassung durch spezialisiertes Personal ist sicher gestellt. Ein entsprechendes Know-how für die Dokumentenerfassung kann aufgebaut werden. Dies ist besonders dann wichtig, wenn das Unternehmen eine Strategie der frühen Erfassung verfolgt und alle Informationen elektronisch den Mitarbeitern zur Verfügung stellen will.
Dezentrales Scannen macht dort Sinn, wo eine verteilte Unternehmensstruktur mit zahlreichen Standorten zu finden und die Dokumente erst nach der Bearbeitung erfasst werden.
Als Sonderfall sind hier Multifunktionsgeräte wie kombinierte Kopierer/Fax/Drucker/Scanner-Systeme zu nennen. Diese ermöglichen unabhängig von spezialisierten zentralen oder dezentralen Scan-Strecken die Erfassung von kleineren Dokumentenbeständen.

„Frühes, paralleles, spätes“ Scannen

Je nach Zeitpunkt des Scannens im Prozess der Erfassung und Sachbearbeitung unterscheidet man in „frühes“, „paralleles“ und „spätes“ Scannen.
Beim „frühen“ Scannen erfolgt die Wandlung des Papierdokuments in ein elektronisches Dokument vor der eigentlichen Bearbeitung. Dazu muss beim Scannen bereits mindestens eine Basisindizierung und eventuell eine elektronische Signatur erfolgen. Die weitere Bearbeitung des dem Dokument zugrunde liegenden Vorgangs kann dann elektronisch erfolgen. Das frühe Scannen ist die Basis für komplett elektronisch unterstützte Workflows, elektronische Aktenlösungen, papierlose Sachbearbeitung u.ä.
Beim „späten“ Scannen erfolgt der Scan-Vorgang erst nach Bearbeitung des Dokuments. Häufig wird bei der Bearbeitung ein Barcode oder eine eindeutig zu identifizierende Nummer auf das Dokument angebracht. Dann kann durch einen Erkennungsvorgang das Dokument nach dem Scannen direkt einer Anwendung zugeordnet und automatisch indiziert werden.
Es sind auch Mischformen des frühen und späten Scannens denkbar. Beim parallelen Scannen wird das Dokument am Arbeitsplatz des Bearbeiters bearbeitet, gescannt und indiziert.

Einbindung von Multifunktionsgeräten

Multifunktionsgeräte wie kombinierte Kopierer/Fax/Drucker/Scanner-Systeme (MFP = Multi Function Printer) können innerhalb eines ECM-Systems als dezentrale Scanner Einsatz finden. Ein MFP versendet automatisch gescannte Dokumente in das ECM-System und dessen Workflow. Die Indexkriterien/Metadaten werden anhand von ausgewählten Profilen (z.B. „Rechnung“) manuell erfasst. Die Vorteile bei der Nutzung

solcher Geräte in einem ECM-System sind, dass es das MFP erlaubt, Papierdokumente in Informationen umzuwandeln, die sich auf einfache Weise in alle Geschäftsanwendungen integrieren lassen und dass bestehende Komponenten genutzt werden können.
Allerdings eignet sich ein MFP nur für kleine Scan-Mengen und für eine einfache Indizierung. Am MFP selbst ist keine Prüfung der Scan-Qualität möglich (beschränkte Displays). Für eine komplexere Indizierung, Workflow-Steuerung und Qualitätsprüfung ist Zusatzsoftware erforderlich. Als Problempunkte sind generell zu nennen:

- Ist eine Schnittstelle zu ECM-Systemen oder Zusatzsoftware vorhanden?
- Wie wird die Lieferung von nicht indizierten Dokumenten in empfangende Systeme (z.B. ECM-System) geregelt?
- Wie wird das Nachscannen bei schlechter Qualität durchgeführt?

Buchscannen
Bücher zu scannen erfordert eine komplett andere technische Vorgehensweise als Dokumentenscannen. Anders als bei herkömmlichen Flachbettscannern liegt das Buch bei den Buchscannern nicht mehr mit seinen Seiten nach unten auf einer Glasplatte. Der Grund liegt verständlicherweise darin, dass dieses zusätzliche „von oben mit einem Deckel beschwert zu sein“ den Buchrücken zu stark belastet und das Buch beschädigt werden könnte. Deswegen liegt das Buch bei den Auflichtscannern meist in einer Buchwippe und die Seiten werden von oben gescannt. Das vermeidet starke Gebrauchspuren nach der Digitalisierung eines Werkes. Da zu scannende Seiten so nicht plan liegen können, werden erhöhte Anforderungen an die Imageverarbeitung bezüglich der Bildqualität gestellt.
Buchscanner sind grundsätzlich in drei verschiedene Kategorien einzuteilen. Manuelle Buchscanner (Auflichtscanner oder adaptierte Flachbettscanner), halbautomatische oder vollautomatische Buchscanner. Während man bei den manuellen Buchscannern die Seiten von Hand umblättert, haben vollautomatische Geräte dafür verschiedene Mechanismen. Es gibt Buchscanner, die bis zu Vorlagen der Größe DIN A0 verarbeiten können.

Sonderformate
An dieser Stelle soll auf zwei besondere Formate bzw. Speicheranforderungen eingegangen werden.
Die Speicherung von Plänen und Karten wird meist von Spezialanwendungen abgedeckt. Diese sind ausgelegt für die Verwaltung und Darstellung von Plänen und Dokumentationen zu technischen Projekten und stellt deswegen Funktionen bereit, die ein DMS nur eingeschränkt abbilden kann, z.B. übersichtliche Projektverwaltung, Archivierung der Pläne und Dokumentationen, detaillierte Ablage aller Plan beschreibenden Daten, komplette Versionsverwaltung, Adressbuch mit zahlreichen Schnittstellen, vordefinierter und ad-hoc-Planversand, nachvollziehbare Dokumentation aller Planbewegungen, konsequente Terminverfolgung, Plankopfdatenaustausch mit CAD-Systemen, leistungsfähiger Viewer zum Betrachten und Drucken, Drucken vordefinierter und frei erstellbarer Listen sowie Anbindung von Plottern bzw. Plot-/Repro-Betrieben.
Deswegen ist es empfehlenswert, die Pläne in diesem Spezialsystem zu belassen. Die Spezial-Funktionen und die Archivierung in Verbindung mit Plänen bleiben dort. Rückgriffe auf Pläne sollen bei Bedarf direkt aus dem DMS durchgeführt werden.

DICOM (Digital Imaging and Communications in Medicine) ist ein weltweiter offener Standard zum Austausch von digitalen Bildern in der Medizin. Er standardisiert sowohl das Format zur Speicherung von Bilddaten, als auch das Kommunikationsprotokoll zum Austausch der Bilder. Fast alle Hersteller medizinisch bild gebender Systeme wie z.B. Digitales Röntgen, Computertomographie oder Sonografie implementieren den DICOM-Standard in ihren Geräten. Dadurch wird im klinischen Umfeld Interoperabilität zwischen medizinischen Systemen verschiedener Hersteller erreicht.

Scan-Outsourcing

Eine mittlerweile sehr gebräuchliche Weise Dokumente zu erfassen, ist die Vergabe dieser Tätigkeit außer Haus. Dies versteht man unter dem Begriff „Outsourcing“. Outsourcing ist eine langfristige und häufig endgültige Vergabe von Leistungen an einen externen Dienstleister, die bisher selbst erstellt wurden oder sonst selbst erstellt werden müssen. Scan-Outsourcing ist die Vergabe von Scan- und Erfassungs-Leistungen an einen externen Dienstleister.

Die auf dem Markt präsenten Dienstleister bieten eine komplette Palette an: Arbeitsvorbereitung (Entklammern, Glätten, Sortieren), Scannen, Indizierung, Abgleich mit dem ERP-System des Kunden (z.B. mit Bestelldaten), Übertragung der Buchungssätze in das ERP-System, Qualitäts- und Vollständigkeitskontrolle und Nachbehandlung (z.B. Vernichtung) der Belege.

Doch ohne ausreichende Vorbereitung wird die Einbeziehung eines externen Dienstleistungsunternehmens nicht erfolgreich sein. Es sind grundlegende Fragen zu klären (z.B. Welches Ziel wird verfolgt? Ist es unternehmenspolitisch durchsetzbar? Ist das Volumen ausreichend? Sind die Prozesse überhaupt abgrenzbar?). Die eigenen Prozesse müssen mit Spezifikation der Schnittstellen, Angaben zur Dokumentenstruktur und zu Benutzergruppen dokumentiert werden. Generelle Anforderungen und geforderte Service Levels müssen festgelegt werden. Eine Wirtschaftlichkeitsbetrachtung sollten durchgeführt werden und Angebote sollten eingeholt und verglichen werden.

Üblicherweise wird Outsourcing an einen Dienstleister vergeben, der die Aufgaben in seinen Räumlichkeiten durchführt. Es gibt aber auch die Variante des Inhouse Outsourcing. In diesem Fall „übernimmt“ der Dienstleister praktisch die Scan-Stelle des Auftraggebers mit seinem Personal oder er bringt sogar seine Scanner mit.

Altakten-Scannen

Bei der Neueinrichtung einer ECM-Lösung stellt sich oft die Frage, wie können „alte“ Bestände übernommen werden. Natürlich können diese eingescannt werden. Da es sich hierbei meist um sehr große Bestände handelt, sollte dies konzeptionell gut durchdacht werden. Grundsätzlich bieten sich folgende Varianten an: der gesamter Altbestand wird gescannt, der Altbestand ab einem bestimmten Stichtag wird gescannt, die Akten werden abhängig von geschätzter Zugriffshäufigkeit gescannt oder sie werden on demand d.h. bei Zugriff gescannt.

Für die Übernahme großer Altaktenbestände sollte stets ein externer Dienstleister herangezogen werden. Es lohnt sich nicht, hierfür die Infrastruktur und Prozesse im Haus bereitzustellen (Hard- und Software, Personal, Raum).

Fazit / Ausblick aktuelle Entwicklungen

Abschließend sollen ein paar Fragen mitgegeben werden. Mal sehen wie sich die Anwender, die Hersteller und der gesamte Markt verhält.

- PDF/A als DAS Langzeitarchivformat?! Viele Capture-Systeme liefern beim Scannen noch TIFF
- Ist Farbe beim Scannen rechtsrelevant?
- Benötigen wir den Begriff des „Elektronischen Originals“ wo in IT-Systemen meistens mit Kopien gearbeitet wird?
- Bleibt es bei nur zwei führenden internationalen Capture-Basistechnologie-Anbietern?
- Wird sich der Markt für automatische Klassifikation noch weiter konsolidieren?
- Wie lange bleibt die Erfassung von Papier noch wichtig, wenn der Trend zu immer mehr elektronisch originär entstandenen Dokumenten geht?
- Nur das Scannen outsourcen oder gleich die ganze Lösung außer Haus geben?

Fachlich kann festgestellt werden:

- Capture“ hat sich zu einer eigenständigen Disziplin entwickelt sowohl für die Belieferung von ECM-Systemen und Archiven als auch für die Lieferung von Daten an operative Systeme.
- Die automatische Klassifikation ist praxistauglich, besonders wenn die Ergebnisse mit vorhandenen Daten abgeglichen werden können.

Autor

Dr. Ulrich Kampffmeyer, Jahrgang 1952, ist Gründer und Geschäftsführer der PROJECT CONSULT Unternehmensberatung GmbH, Hamburg, eine der führenden produkt- und herstellerunabhängigen Beratungsgesellschaften für ECM Enterprise Content Management, BPM Business Process Management, Knowledge Management, Records Management, Collaboration, Archivierung, Enterprise 2.0 und Information Management.

Er beriet namhafte Kunden aller Branchen im In- und Ausland bei der Konzeption und Einführung von ECM-Lösungen.

Als Gründer und langjähriger Vorstandsvorsitzender nationaler und internationaler Branchenverbände prägte er wesentlich den deutschen Markt für ECM. Dr. Kampffmeyer ist Mitglied in mehreren internationalen Standardisierungsgremien im Umfeld des Workflow-, Dokumenten- und Records-Management.

Dr. Kampffmeyer ist anerkannter Kongressleiter, Referent und Moderator zu Themen wie elektronische Archivierung, Records Management, Dokumentenmanagement, Workflow, Rechtsfragen, Business Re-Engineering, Wissensmanagement und Projektmanagement. Auf zahlreichen nationalen und internationalen Kongressen und Konferenzen wirkte er als Keynote-Sprecher mit. Er wurde mehrfach von der ComputerWoche zu den 100 wichtigsten Persönlichkeiten der deutschen IT-Branche gezählt. Von internationalen Verbänden erhielt er zahlreiche Auszeichnungen für sein Wirken als „ECM Mentor“ in Europa.

Autorenrecht und CopyRight
Autor: Dr. Ulrich Kampffmeyer
PROJECT CONSULT Unternehmensberatung GmbH
Breitenfelder Str. 17
D-20251 Hamburg
Tel.: 040 / 460 762 20
Fax: 040 / 460 762 29
E-Mail: Presse@PROJECT-CONSULT.com
Web: www.PROJECT-CONSULT.com